Die relative Verteilung als Ansatz zur Analyse von Gruppenunterschieden

Ben Jann

ETH Zürich, jannb@ethz.ch

Ludwig-Maximilians-Universität München
23. Juni 2009

Gliederung

Einleitung

- Ziel: Vergleich von zwei Gruppen (oder Zeitpunkten) bezüglich eines kontinuierlichen Merkmals.
- Ein prominentes Beispiel ist die Analyse von Erwerbseinkommen bzw. Löhnen nach Geschlecht.
- Aus Gründen der Einfachheit werden solche Vergleiche häufig nur anhand einiger weniger, als zentral angesehener Masszahlen durchgeführt (i.d.R. Erwartungswert).

Einleitung

- Beispiel: Einkommen/Löhne und Geschlecht.

  Wie wird das analysiert?

  - öffentliche Statistik: Differenz im Mittelwert (oder Median) der (standardisierten) Löhne
  - Kontrolle von Drittvariablen I: Geschlecht als Dummy-Variable in einem Regressionsmodell $\Rightarrow$ konditionale Mittelwertsdifferenz
  - Kontrolle von Drittvariablen II (kontrafaktischer Ansatz): Dekomposition der (logarithmierten) Lohnunterschiede in einen „erklärten Teil“ (Effekt der Unterschiede in den Drittvariablen) und einen „unerklärten Teil“ (Effekt der Unterschiede in den Koeffizienten; Diskriminierung?) (Blinder 1973, Oaxaca 1973, etc.)

Einleitung

- Solche Analysen sind zwar informativ, decken aber nicht immer alle wichtigen Aspekte ab.
- Wünschenswerte sind deshalb (nicht-parametrische) Verfahren, mit denen Verteilungen detailliert verglichen werden können.

Einleitung

- Einige Ansätze:
  - Semi-parametrische Erweiterung der Blinder-Oaxaca-Dekomposition auf beliebige Masszahlen (Quantile, Streuung, etc.) mit Hilfe der Invertierung der Verteilung von Residuen aus Regressionsmodellen (Juhn, Murphy und Pierce 1993; Blau und Kahn 1996a).
  - Mit einem ähnlichen Ansatz: Analyse der Veränderung von Gruppenunterschieden unter Berücksichtigung der „allgemeinen“ Veränderung der Verteilung (Juhn, Murphy und Pierce 1991; Blau und Kahn 1992, 1996b, 1997).
  - Untersuchung von Verteilungen mit Hilfe von Quantils-Regressionen (Buchinsky 1998); Erweiterung der Blinder-Oaxaca-Dekomposition auf Quantile (Machado und Mata 2005); nicht-parametrische Blinder-Oaxaca-Dekomposition mit Hilfe von Matching (Ñopo 2004).
  - Analyse von Differenzen in Dichtefunktionen; kontrafaktische Betrachtung mit Hilfe von Gewichten (DiNardo, Fortin und Lemieux 1996).

Einleitung

- Die relative Verteilung: Weiterer (nicht-parametrischer) Ansatz zur Visualisierung und Analyse der Unterschiede oder Veränderungen von Verteilungen.
- Einige zentrale Literaturhinweise: Morris, Bernhardt und Handcock (1994), Bernhardt, Morris und Handcock (1995), Handcock und Morris (1998, 1999), Handcock und Janssen (2002).
- Grundlegender Gedanke: Interpretation der Werte von Gruppe A als relative Positionen in der Verteilung von Gruppe B $\Rightarrow$ Analyse der Verteilung von „relativen Rängen“.
- Eine bemerkenswerte Eigenschaft des Ansatzes ist, dass die Resultate weitgehend unabhängig sind von monotonen Transformationen der Daten (z.B. Löhne versus logarithmierte Löhne).
- Der Ansatz ist eng verwandt mit dem Ansatz von DiNardo, Fortin und Lemieux (1996).

Relative Daten: Definition

- Sei Y_0 das interessierende Merkmal in der Referenzgruppe und Y das Merkmal in der Vergleichsgruppe. Die dazugehörigen Dichtefunktionen (PDF) bzw. kumulativen Verteilungsfunktionen (CDF) werden mit $f_0(y)$ und $f(y)$ bzw. $F_0(y)$ und $F(y)$ symbolisiert.

- Die „relativen Daten“ (relativen Ränge) sind dann definiert als

$$R = F_0(Y), \qquad R \in [0, 1]$$

  Das heisst, man erhält die relativen Daten, indem man die Verteilungsfunktion der Referenzgruppe auf die Daten der Vergleichsgruppe anwendet.

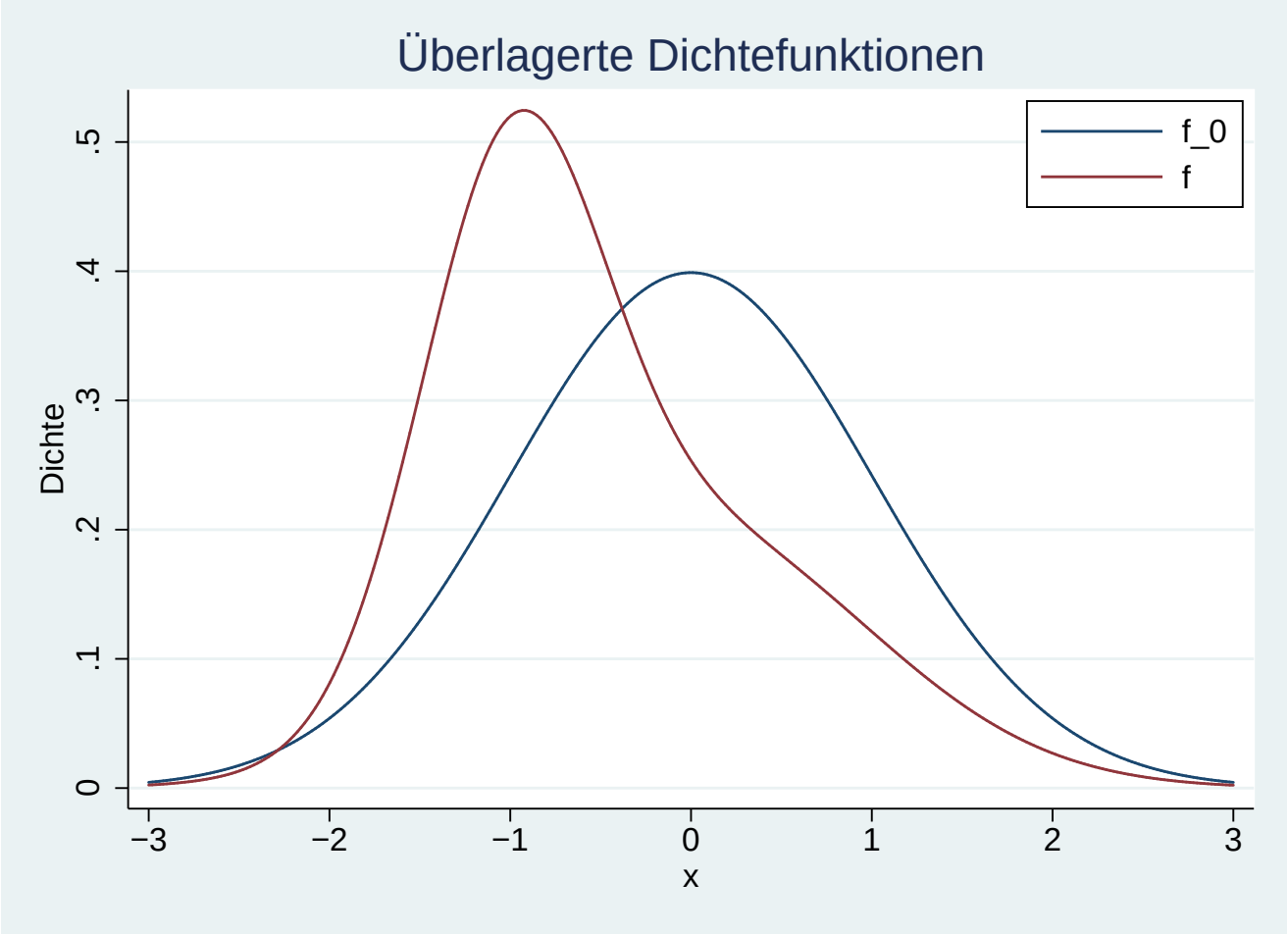
Relative Verteilungsfunktion

- Die kumulative Verteilungsfunktion (CDF) der relativen Daten R ist dann gegeben als

$$G(r) = F(F_0^{-1}(r)), \qquad 0 \leq r \leq 1$$

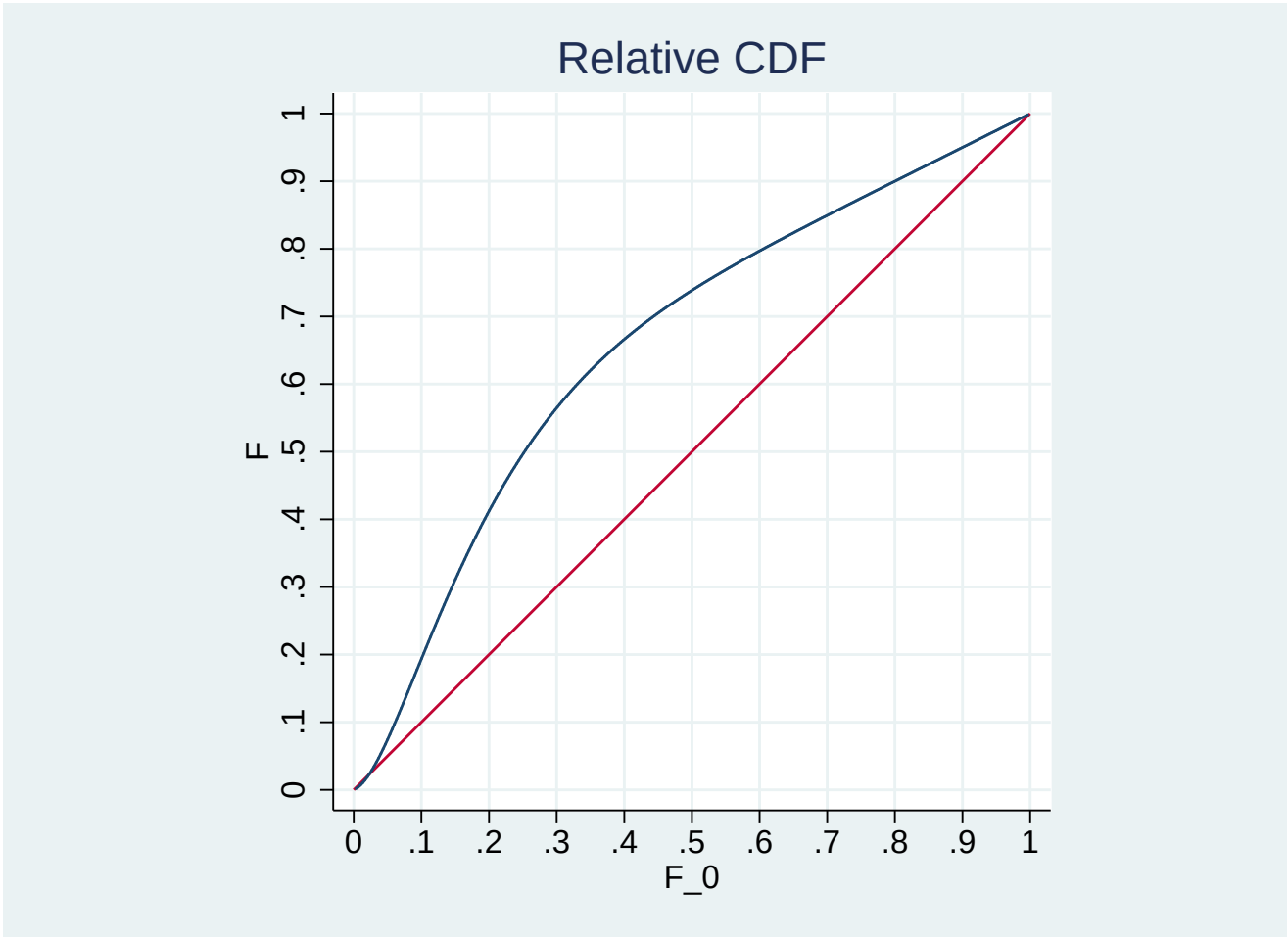
wobei F^{-1} die Inverse von F, also die Quantils-Funktion symbolisiert.

Veranschaulichung: Dichtefunktion für zwei Gruppen




```
two    fun normalden(x)                                , range(-3 3)      ///
    || fun 1/2*normalden(x) + 1/2*normalden(x,-1,1/2), range(-3 3)      ///
    || , xlab(-3(1)3) yti("Dichte") ti("Überlagerte Dichtefunktionen")  ///
         legend(order(1 "f_0" 2 "f") pos(2) ring(0) col(1)) name(a)
```

Relative Verteilungsfunktion (P-P plot)




```
two pci 0 0 1 1 , lsty(yxline)                    ///
 || fun (1/2*normal(invnormal(x))                 ///
        + 1/2*normal((invnormal(x)+1)/0.5))       ///
    , psty(p1) legend(off) xlabel(0(.1)1,grid)    ///
    ylabel(0(.1)1,grid) ti("Relative CDF")        ///
    xti("F_0") yti("F") aspectratio(1)
```

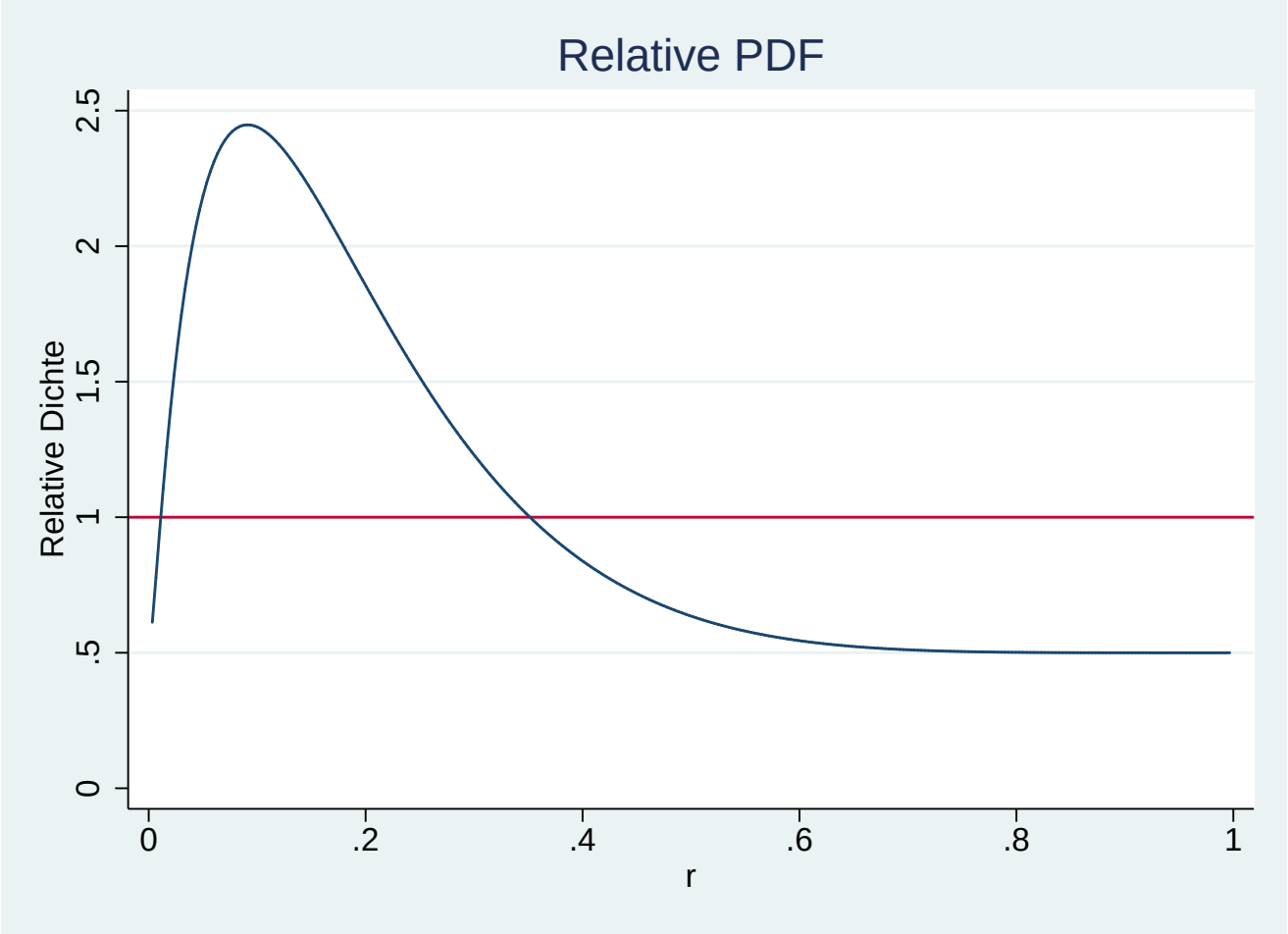
Relative Dichte

- Anschaulicher als die relative Verteilungsfunktion ist die „relative Dichte“.
- Die relative Dichte entspricht der Dichte der relativen Daten R und ist gegeben als

$$g(r) = \frac{f(F_0^{-1}(r))}{f_0(F_0^{-1}(r))}, \qquad 0 \leq r \leq 1$$

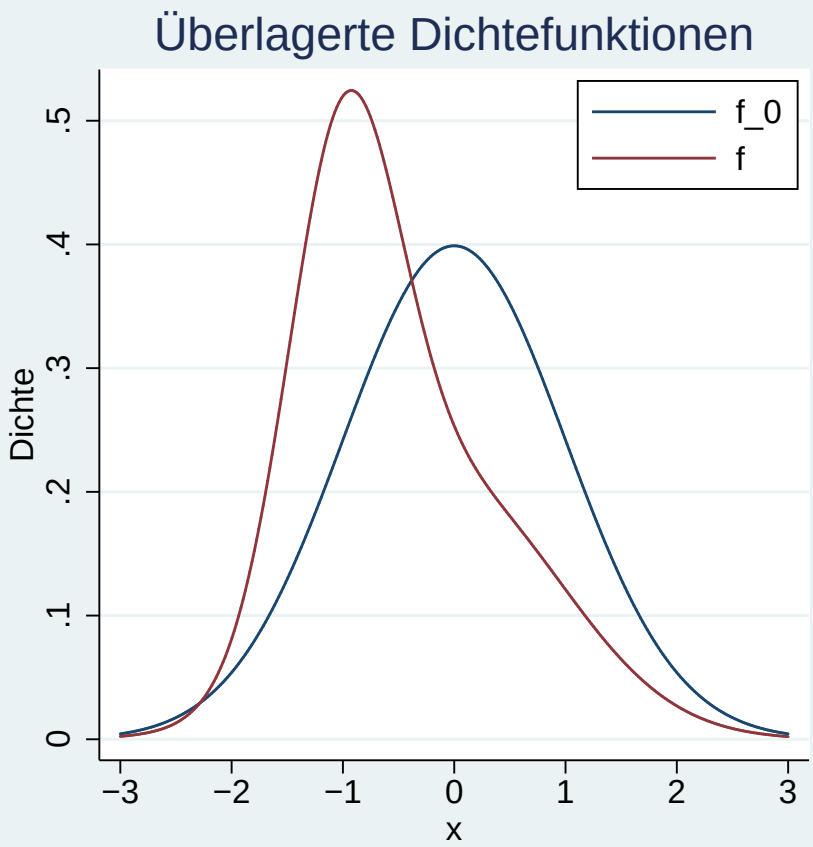
- Die relative Dichte entspricht also dem Verhältnis der Dichten der beiden Gruppen, evaluiert an den Quantilen der Referenzgruppe.
- Die relative Dichte ist eine echte Dichte, d.h. sie integriert zu 1.
- R folgt einer Gleichverteilung (relative Dichte gleich 1), falls es zwischen den Verteilungen der beiden Gruppen keine Unterschiede gibt.

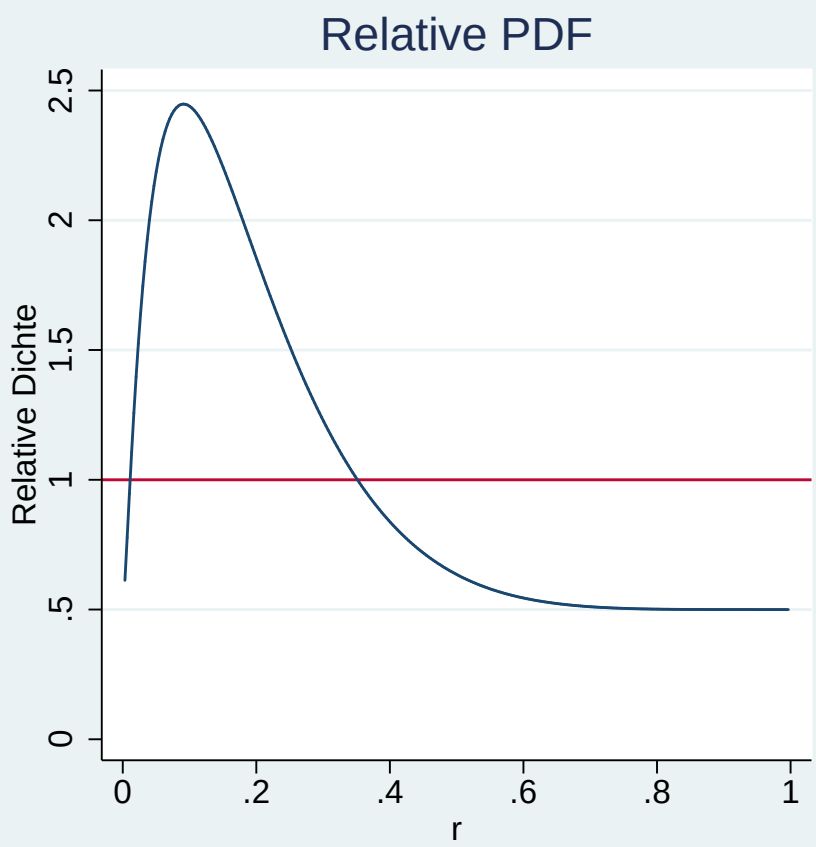
Veranschaulichung: Relative Dichte




```
two fun (1/2*normalden(invnormal(x))                  ///
        + 1/2*normalden(invnormal(x),-1,1/2))         ///
        / normalden(invnormal(x))                     ///
        , yline(1) ti("Relative PDF")                 ///
        ylabel(0, add) yti("Relative Dichte") xti("r") name(b)
```

Veranschaulichung: Relative Dichte






```
graph combine a b, xsize(7.5) iscale(1)
```

Dekomposition von Lage- und Formunterschieden

- Unterschiede in der Verteilungsform werden dann sichtbar, wenn die Lage der Verteilungen angeglichen wird.
- Dekomposition von Lage- und Formunterschieden:

$$\frac{f(y_r)}{f_0(y_r)} = \frac{f_A(y_r)}{f_0(y_r)} \times \frac{f(y_r)}{f_A(y_r)}$$

$$\text{Total} \;=\; \text{Lage} \;\times\; \text{Form}$$

  wobei $y_r = F_0^{-1}(r)$, $r \in [0, 1]$.
- $F_A(y)$ ist eine Verteilungsfunktion mit angepasster Lage. Zum Beispiel:

$$F_A(y) = F_0(y + \rho)$$

  wobei

$$\rho = \text{Median}(Y) - \text{Median}(Y_0)$$

- Alternativ könnte auch das arithmetische Mittel verwendet werden. Je nach Art der Daten kann zudem eine multiplikative Transformation sinnvoll sein.

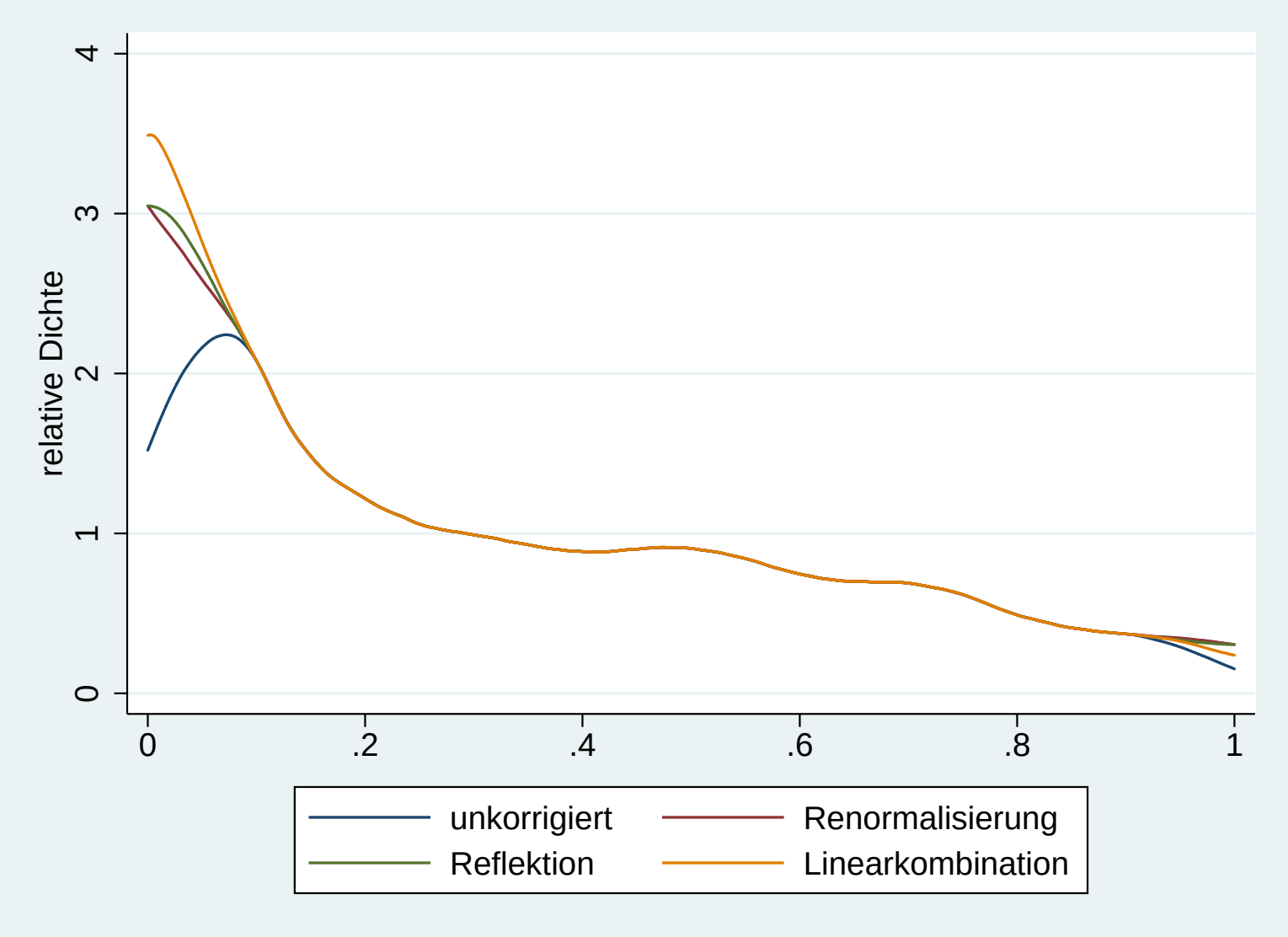
Kontrolle von Drittvariablen

- Die kontrafaktische Verteilungen unter Kontrolle einer Drittvariable X kann ganz einfach durch Gewichtung mit der relativen Dichte von X simuliert werden.
- Bei mehreren Kontrollvariablen ist dies aufgrund der Multidimensionalität nicht mehr möglich. Eine Lösung ist die Verwendung von Gewichten, die aus der Modellierung der Gruppenzugehörigkeit abgeleitet werden (propensity-score reweighting).
- Die Gewichte können allgemein auch mit Matching-Methoden ermittelt werden.
- Grundsätzliches Problem: Die individuellen Beiträge einzelner Drittvariablen können nur sequenziell bestimmt werden (Pfadabhängigkeit).

Schätzung der relativen Dichte: einige Komplikationen

- Relative Daten liegen zwischen null und eins. Übliche Kerndichte-Schätzer sind in diesem Fall ungeeignet, da an den Rändern starke Verzerrungen (nach unten) entstehen. Es müssten also entsprechend korrigierte Schätzer verwendet werden.
- Die Resultate von Dichteschätzungen hängen vom Grad der Glättung ab. Verschiedene Ansätze zur Bestimmung der optimalen Glättung für Kerndichte-Schätzer existieren. Für relative Daten werden allerdings einige Anpassungen benötigt (vgl. z.B. Cwik and Mielniczuk 1993).
- Statistische Inferenz für relative Daten? Die Schätzung der Varianzen ist nicht ganz trivial und approximative Standardformeln sind nicht besonders präzise für endlichen Stichproben. Replikationstechniken (Bootstrap, Jackknife) können aber einfach angewendet werden.

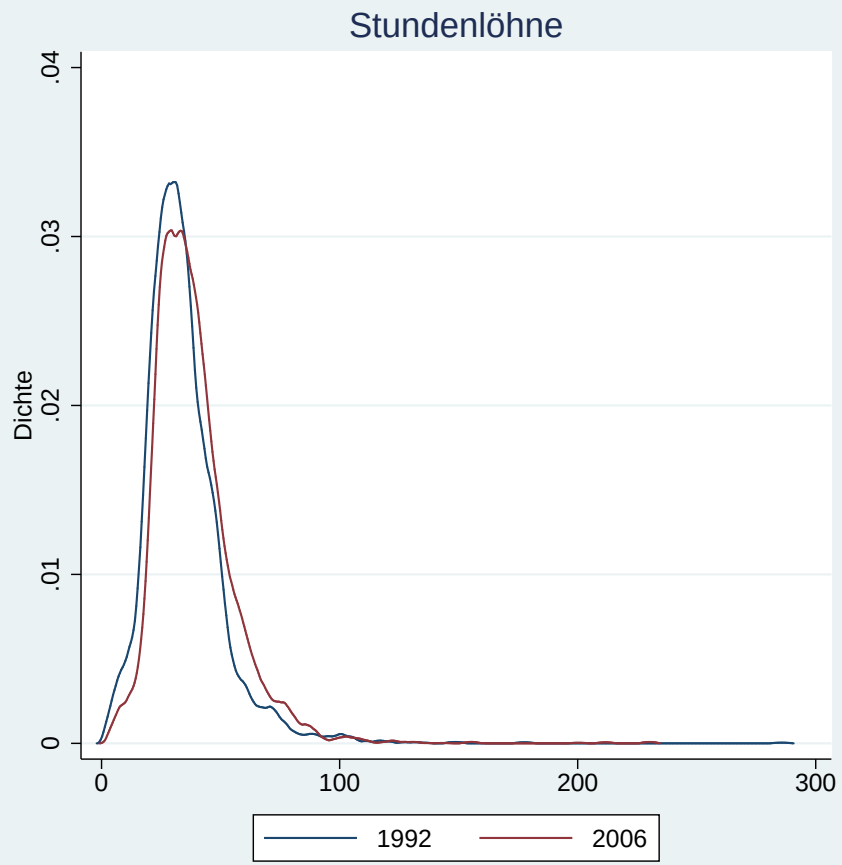
Grenzkorrektur bei der Schätzung der relativen Dichte

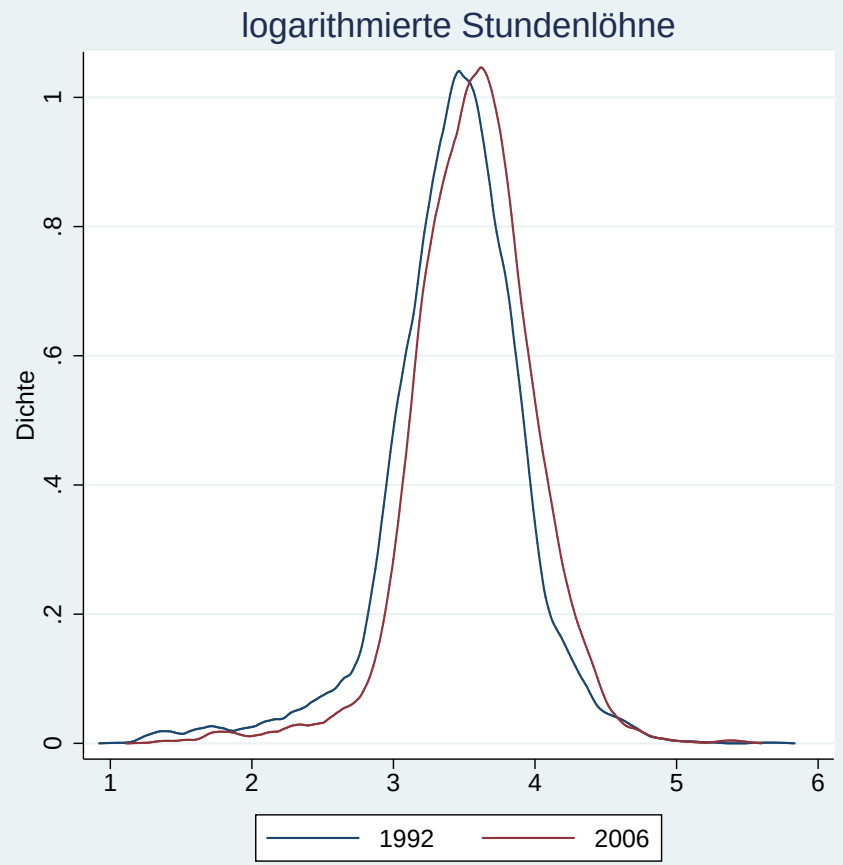


Anwendungsbeispiele

- Daten: Schweizerische Arbeitskräfteerhebung (SAKE) 1991–2006 des Bundesamts für Statistik
- Vergleich der Stundenlöhne von Frauen über die Zeit
- Vergleich von Stundenlöhnen nach Geschlecht
- Auswahl
  - Alter 20–62
  - nur Arbeitnehmerinnen/Arbeitnehmer
  - Arbeitszeit $\geq$ 6 Stunden/Woche
  - nur Schweizerinnen/Schweizer

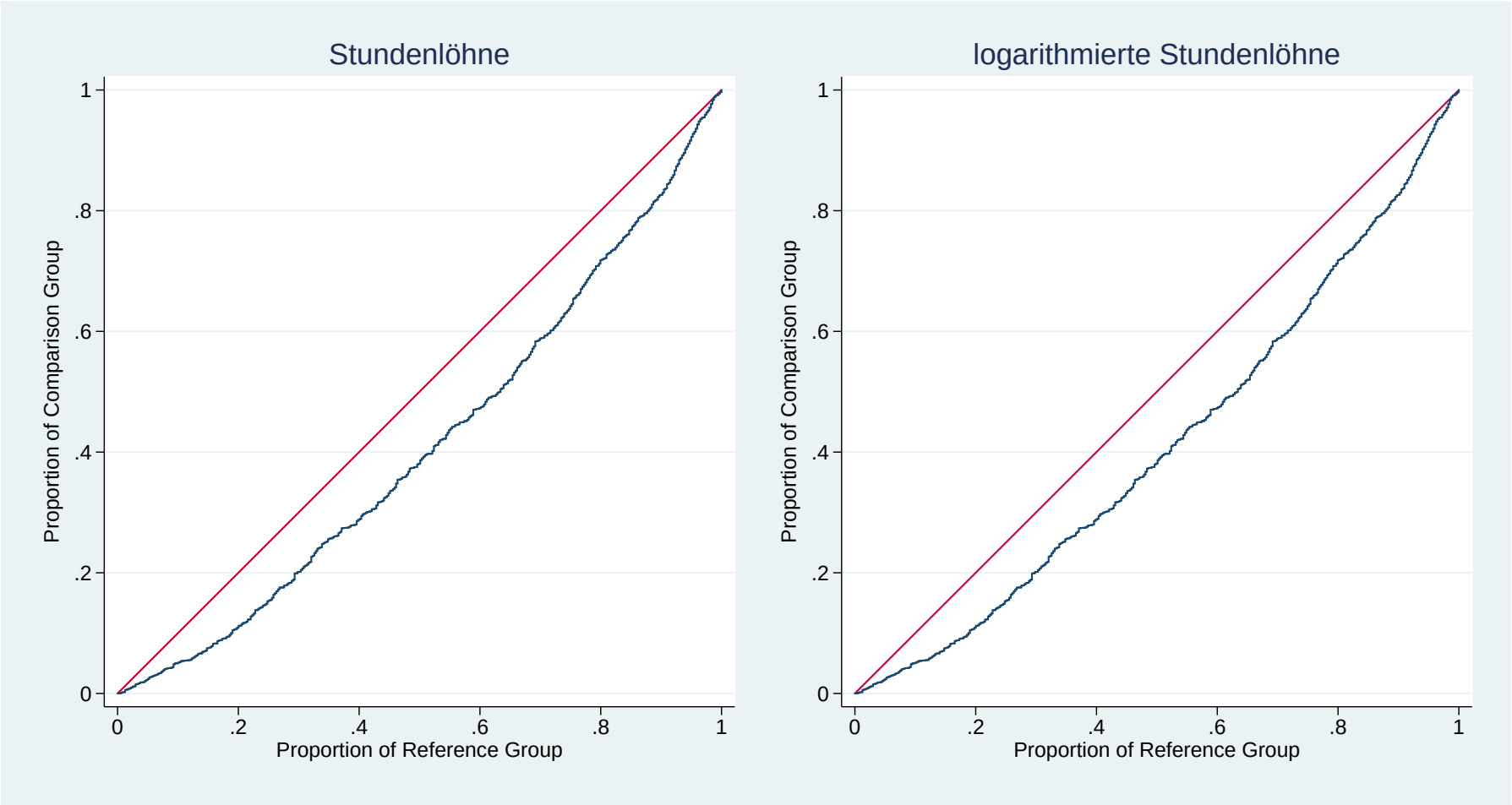
Stundenlöhne von Frauen 1992-2006: Dichte





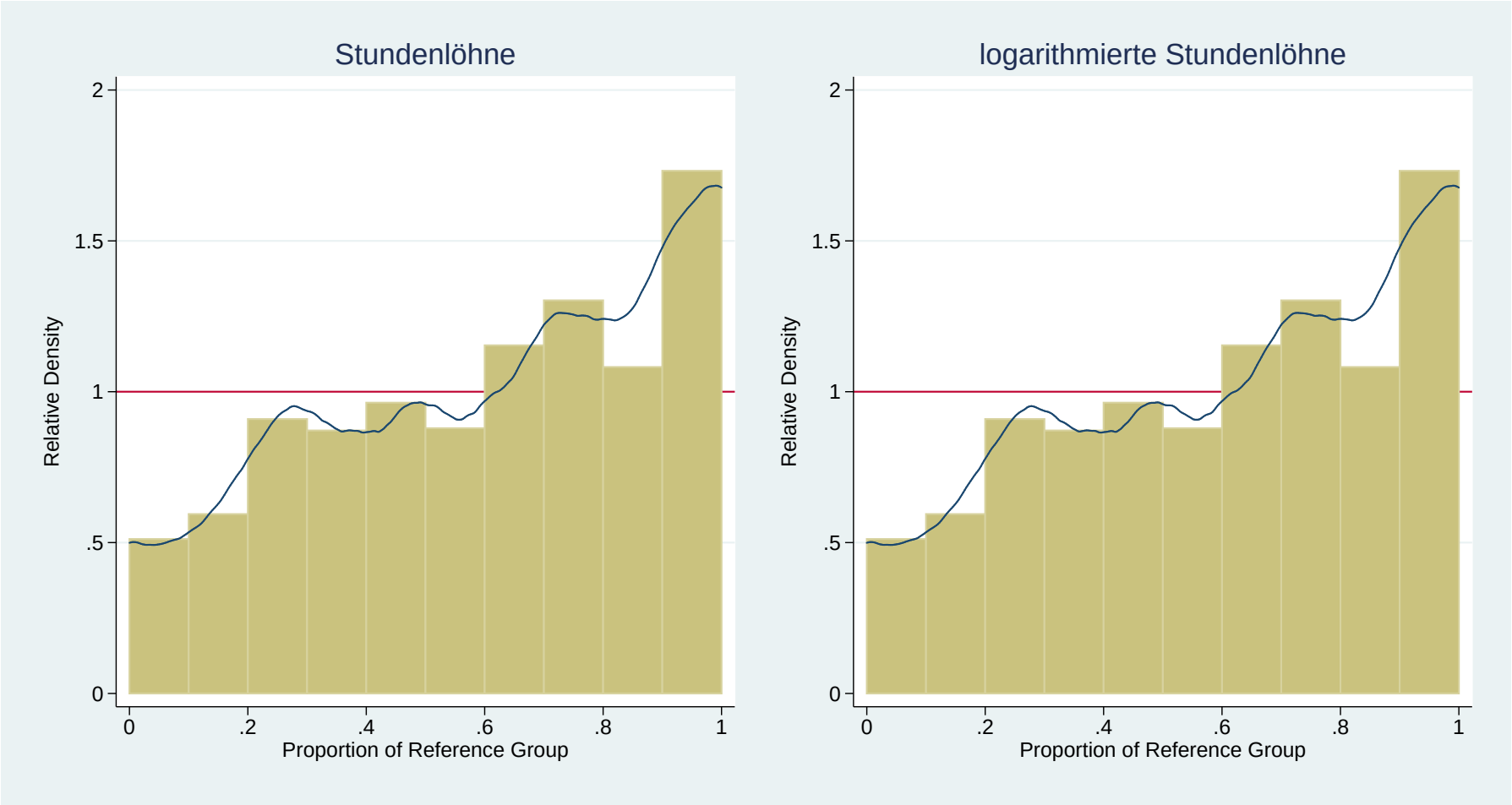

```
. use reldist, clear
(Excerpt from the Swiss Labor Force Survey (SLFS) 1991 - 2006)

. two kdens wage if year==1992 & inlist(female,1) [pw=wt], bw(sj) ///
>  || kdens wage if year==2006 & inlist(female,1) [pw=wt], bw(sj) ///
>  ti(Stundenlöhne) yti(Dichte) xti("") name(a) legend(order(1 "1992" 2 "2006")
> )
(bandwidth = 5.0365478)
(bandwidth = 4.4375098)
```

Stundenlöhne von Frauen 1992-2006: relative CDF



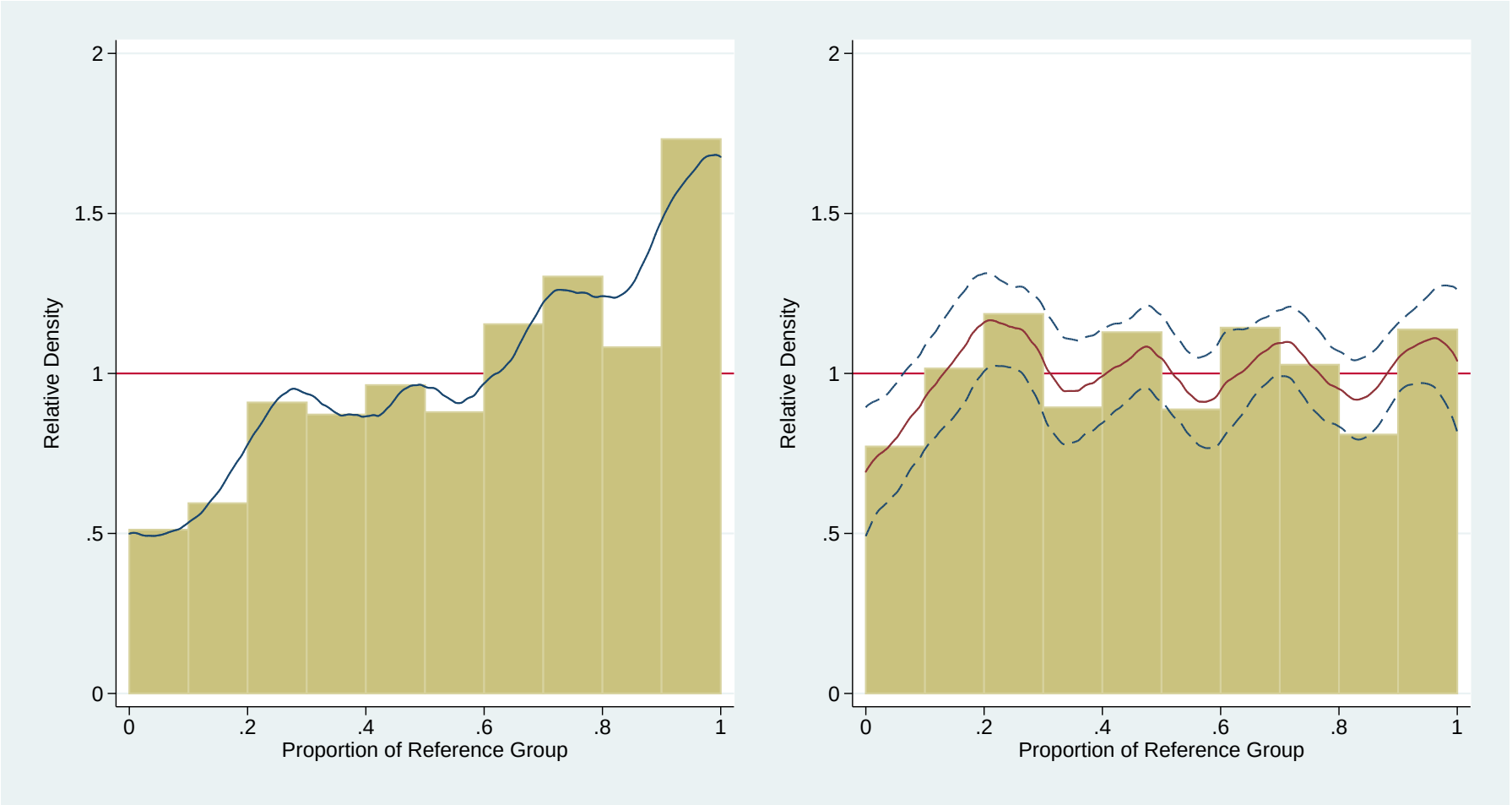

```
. reldist wage if female==1 [pw=wt], by(y0692) cdf ti(Stundenlöhne) name(a)
(reference group: y0692 = 0; comparison group: y0692 = 1)
. reldist lnwage if female==1 [pw=wt], by(y0692) cdf ti(logarithmierte Stundenl
> öhne) name(b)
(reference group: y0692 = 0; comparison group: y0692 = 1)
. graph combine a b, xsize(7.5)
```

Stundenlöhne von Frauen 1992-2006: relative Dichte



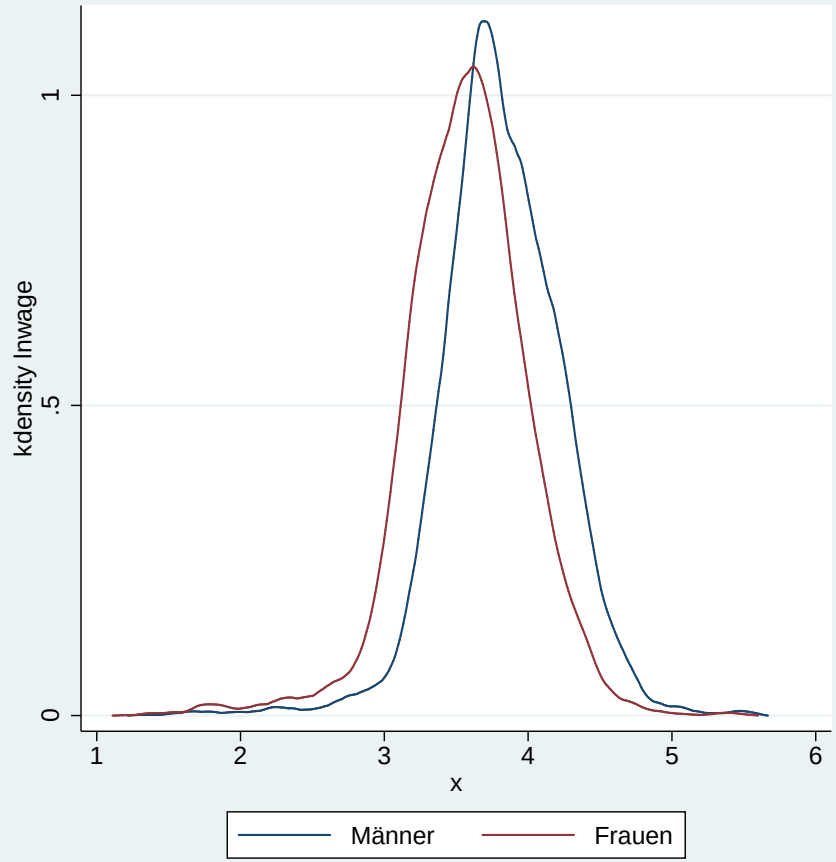

```
. reldist wage if female==1 [pw=wt], by(y0692) bw(sj) pdf hist ti(Stundenlöhne)
>  name(a)
(reference group: y0692 = 0; comparison group: y0692 = 1)
(bandwidth = .101835217)

. reldist lnwage if female==1 [pw=wt], by(y0692) bw(sj) pdf hist ti(logarithmie
> rte Stundenlöhne) name(b)
(reference group: y0692 = 0; comparison group: y0692 = 1)
(bandwidth = .101835217)
```

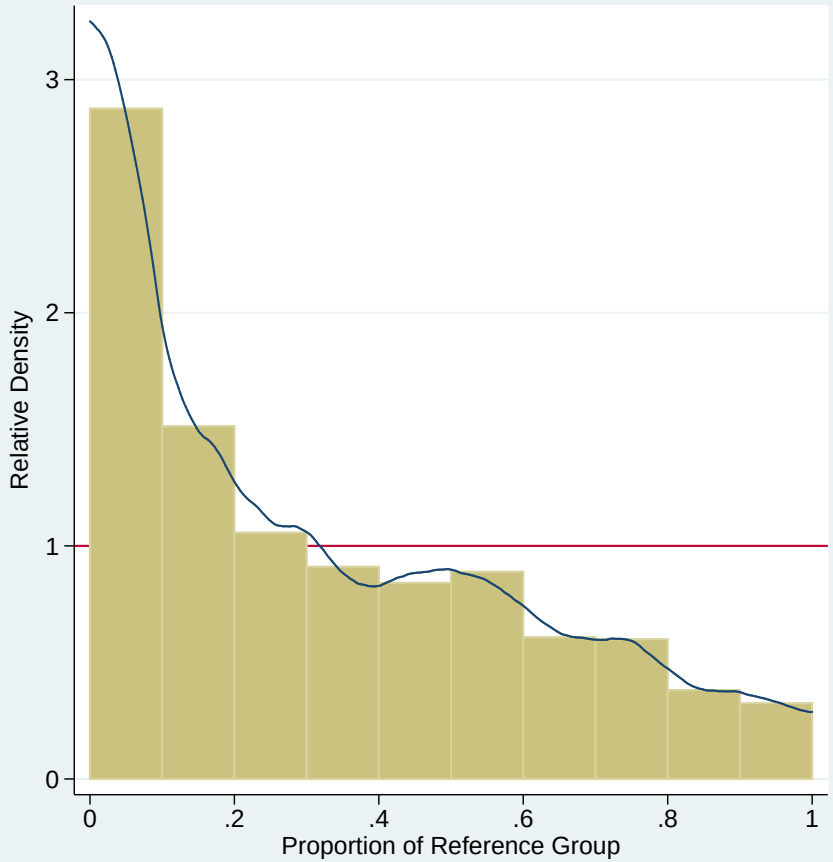
Stundenlöhne von Frauen 1992-2006: Formeffekt




```
. reldist wage [pw=wt] if female==1, by(y0692) bw(sj) ci ///
>     shape multiplicative pdf hist vce(boot, reps(100)) name(a)
(reference group: y0692 = 0; comparison group: y0692 = 1)
(bandwidth = .094174664)

Bootstrap replications (100)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
.................................................   50
.................................................   100
```

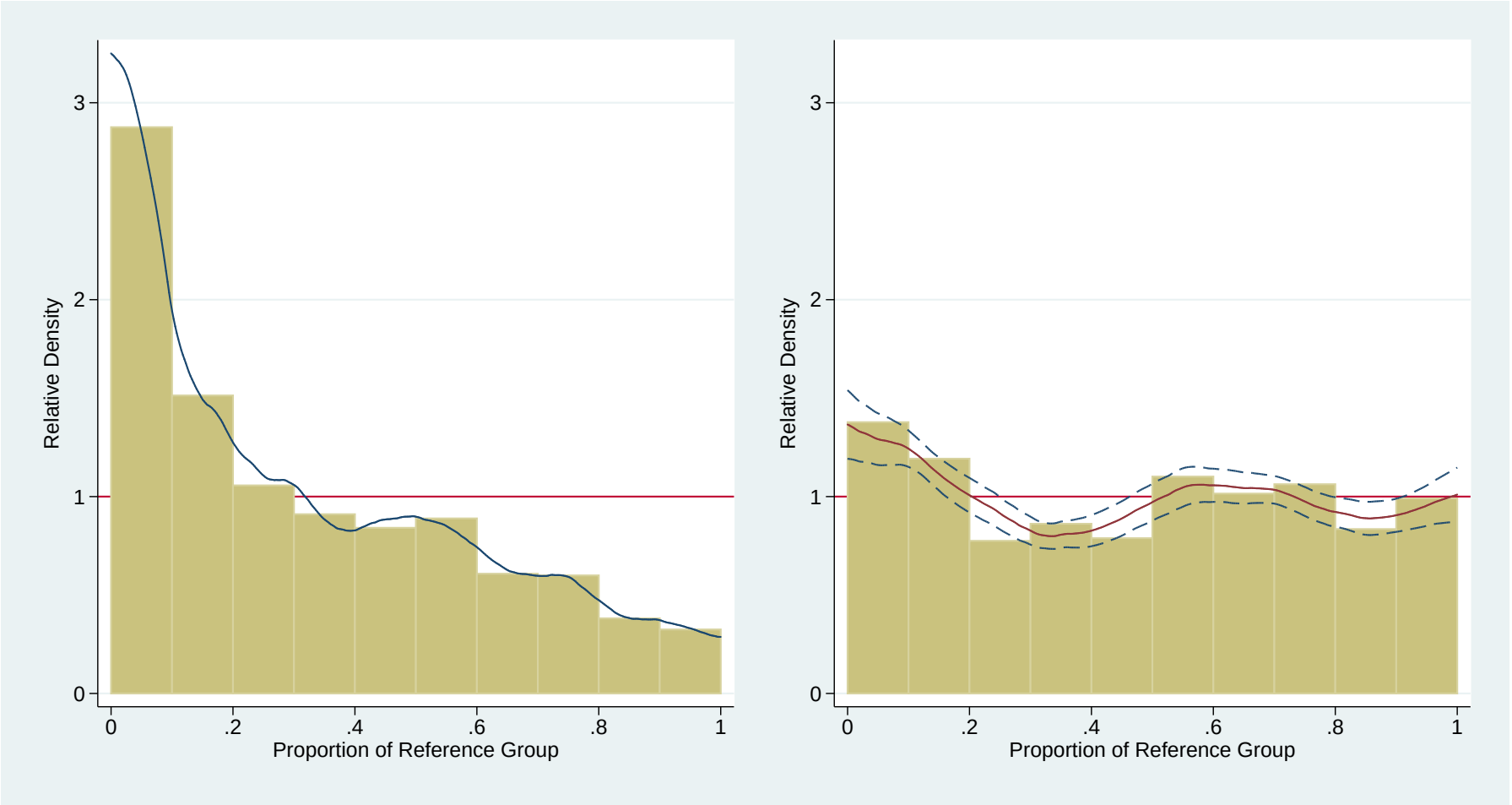
Löhne von Frauen und Männern 2006





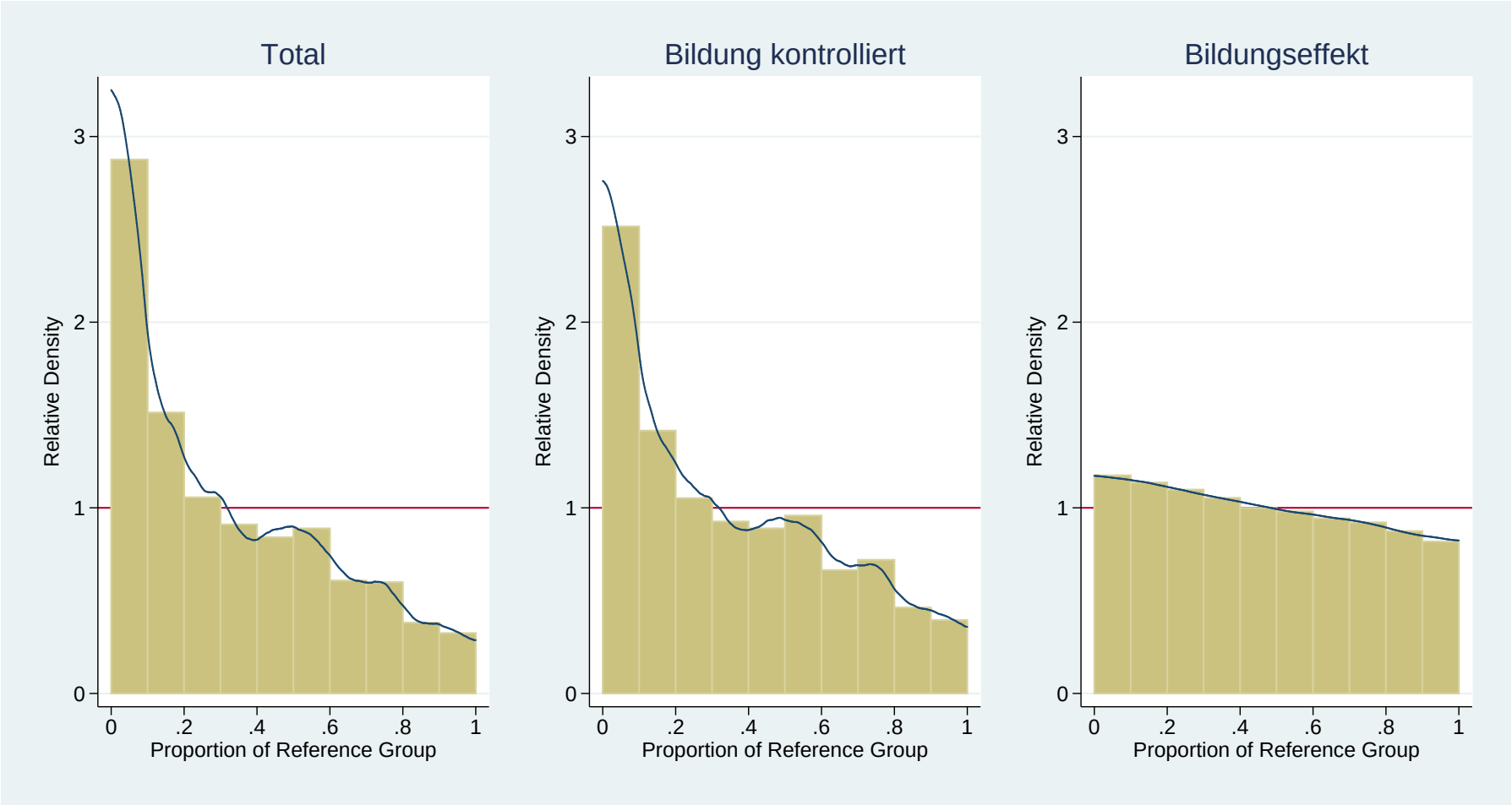

```
. two kdens lnwage if female==0 [pw=wt], bw(sj) ///
>  || kdens lnwage if female==1 [pw=wt], bw(sj) ///
>  legend(order(1 "Männer" 2 "Frauen")) name(a)
(bandwidth = .13849496)
(bandwidth = .16139899)

. reldist lnwage [pw=wt], by(female) bw(sj) pdf hist name(b)
(reference group: female = 0; comparison group: female = 1)
(bandwidth = .070460862)
```

Löhne von Frauen und Männern 2006: Formeffekt



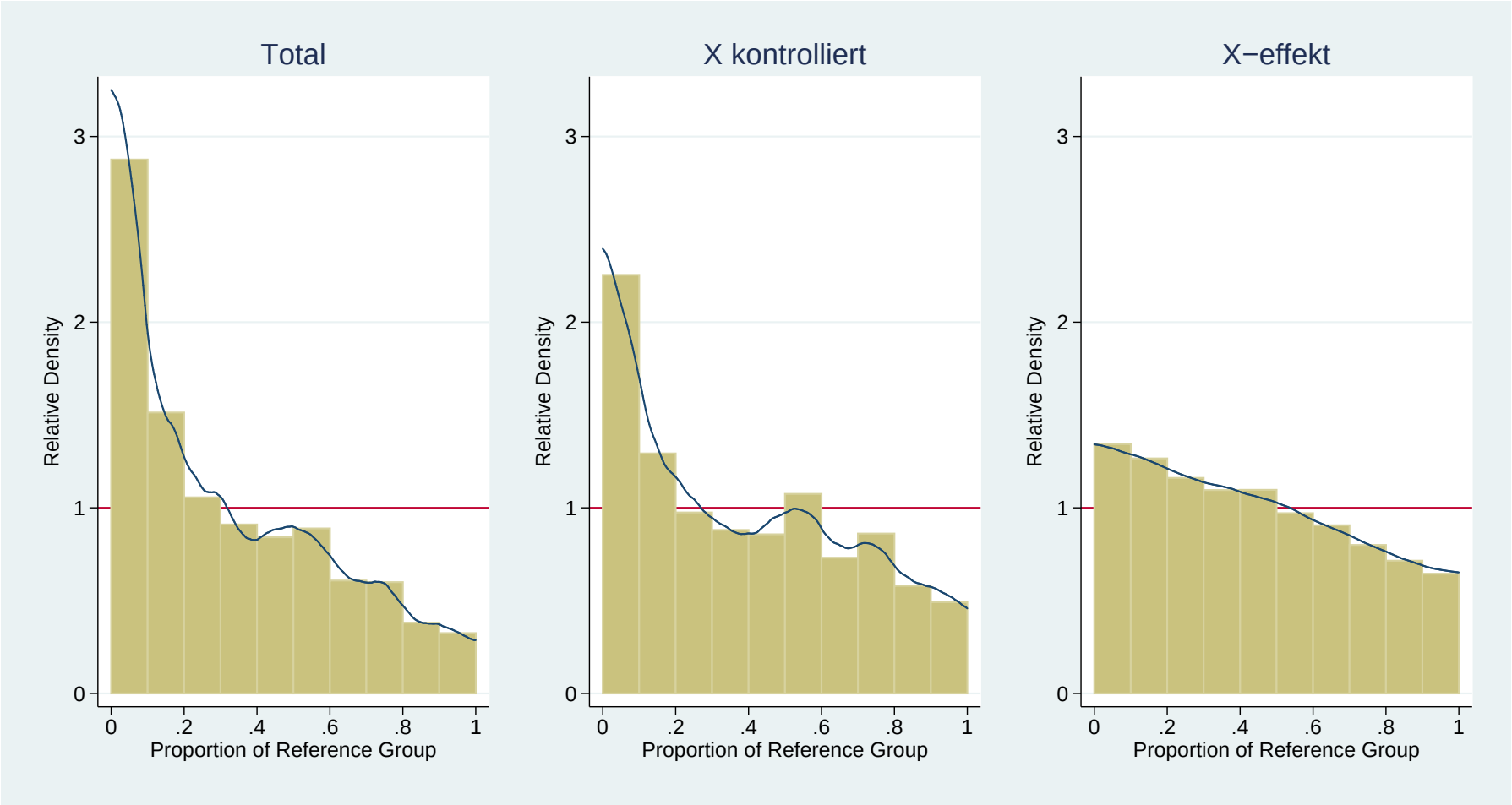

```
. reldist lnwage [pw=wt], by(female) bw(sj) ci ///
>     shape multiplicative pdf hist vce(boot, reps(100)) name(a)
(reference group: female = 0; comparison group: female = 1)
(bandwidth = .131032266)

Bootstrap replications (100)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
.................................................    50
.................................................   100
```

Löhne von Frauen und Männern 2006: Bildungseffekt



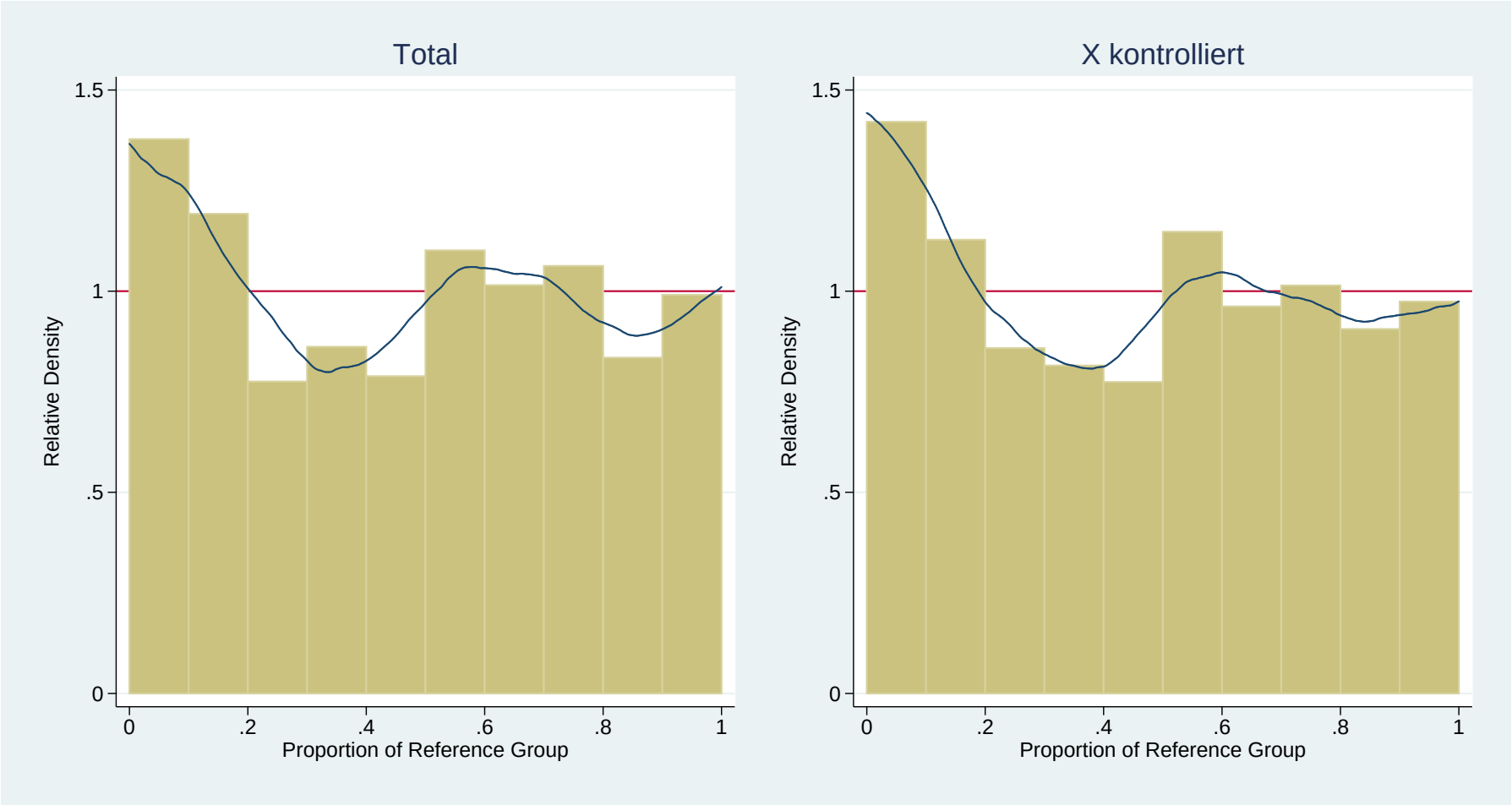

```
. bys year educ: egen sumwtmale = total(wt*(1-female))
. bys year educ: egen sumwtfemale = total(wt*female)
. generate relwt = cond(female, wt * sumwtmale / sumwtfemale , wt)
. expand 2
(9333 observations created)
. bys id: gen byte second = _n==2
```

Effekt von Bildung, Berufserfahrung und Firmentreue




```
. gen exp2 = exp^2
. gen ten2 = tenure^2
. xi: probit female i.educ exp exp2 tenure ten2 [pw=wt]
i.educ            _Ieduc_1-9          (_Ieduc_1 for educ==8 omitted)
(sum of wgt is   1.7939e+06)
Iteration 0:   log pseudolikelihood =  -6427.984
Iteration 1:   log pseudolikelihood = -5983.4641
```

Formeffekt unter Kontrolle von Bildung etc.




```
. reldist lnwage if second==0 [pw=wt], by(female) bw(sj) hist pdf ///
>     shape multiplicative ti(Total) name(a)
(reference group: female = 0; comparison group: female = 1)
(bandwidth = .131032266)

. reldist lnwage [pw=relwt], by(femA0) bw(sj) hist pdf ///
>     shape multiplicative ti(X kontrolliert) name(b)
(reference group: femA0 = 0; comparison group: femA0 = 1)
(bandwidth = .134710224)
```

Zusammenfassung

- Die relative Verteilung erscheint als ein nützliches Konzept zur Analyse von Verteilungsunterschieden. Der Ansatz sollte weiterverfolgt werden.
- Lieder sind in der (sozialwissenschaftlichen) Literatur aber bisher kaum Anwendungen zu finden (mit Ausnahme der zitierten Arbeiten).
- Ein systematischer Vergleich mit anderen Ansätzen, wie z.B. der Dekomposition auf Grundlage von Quantilsregressionen oder dem Ansatz von DiNardo, Fortin und Lemieux (1996) wäre wünschenswert.

Vielen Dank für Ihre Aufmerksamkeit!

Literaturhinweise I

- Bernhardt, Annette, Martina Morris, and Mark S. Handcock (1995). Women's Gains or Men's Losses? A Closer Look at the Shrinking Gender Gap in Earnings. American Journal of Sociology 101(2): 302-328.
- Blau, Francine D., and Lawrence M. Kahn (1992). The Gender Earnings Gap: Learning from International Comparisons. American Economic Review 82(2): 533-538.
- Blau, Francine D., and Lawrence M. Kahn (1996). International Differences in Male Wage Inequality: Institutions versus Market Forces. Journal of Political Economy 104(4): 791-837.
- Blau, Francine D., and Lawrence M. Kahn (1996). Wage Structure and Gender Earnings Differentials: an International Comparison. Economica 63(250): S29-S62.
- Blau, Francine D., and Lawrence M. Kahn (1997). Swimming Upstream: Trends in the Gender Wage Differential in the 1980s. Journal of Labor Economics 15(1): 1-42.
- Blinder, Alan S. (1973). Wage Discrimination: Reduced Form and Structural Estimates. The Journal of Human Resources 8(4): 436-455.

Literaturhinweise II

- Buchinsky, Moshe (1998). The Dynamics of Changes in the Female Wage Distribution in the USA: A Quantile Regression Approach. Journal of Applied Econometrics 13(1): 1-30.
- Cwik, Jan, and Jan Mielniczuk (1993). Data-dependent bandwidth choice for a grade density kernel estimate. Statistics & Probability Letters 16: 397-405.
- DiNardo, John E., Nicole Fortin, and Thomas Lemieux (1996). Labour Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach. Econometrica 64(5): 1001-1046.
- Handcock, Mark S., and Paul L. Janssen (2002). Statistical Inference for the Relative Density. Sociological Methods and Research 30(3): 394-424.
- Handcock, Mark S., and Martina Morris (1998). Relative Distribution Methods. Sociological Methodology 28: 53-97.
- Handcock, Mark S., and Martina Morris (1999). Relative Distribution Methods in the Social Sciences. New York: Springer.
- Juhn, Chinhui, Kevin M. Murphy, and Brooks Pierce (1991). Accounting for the Slowdown in Black-White Wage Convergence. P. 107-143 in: Marvin Kosters (ed.). Workers and Their Wages. Washington, DC: AEI Press.

Literaturhinweise III

- Juhn, Chinhui, Kevin M. Murphy, and Brooks Pierce (1993). Wage Inequality and the Rise in Returns to Skill. Journal of Political Economy 101(3): 410-442.
- Lemieux, Thomas (2002). Decomposing changes in wage distributions: a unified approach. Canadian Journal of Economics 35(4): 646-688.
- Machado, José A. F., and José Mata (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. Journal of Applied Econometrics 20(4): 445-465.
- Morris, Martina, Annette D. Bernhardt, and Mark S. Handcock (1994). Economic Inequality: New Methods for New Trends. American Sociological Review 59(2): 205-219.
- Ñopo, Hugo (2004). Matching as a Tool to Decompose Wage Gaps. IZA Discussion Paper No. 981.
- Oaxaca, Ronald (1973). Male-Female Wage Differentials in Urban Labor Markets. International Economic Review 14(3): 693-709.